Speaker Normalization Based on Test to Reference Speaker Mapping

Marcel Ogner and Zdravko Kacic

University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, 2000 Maribor, SLOVENIA
marcel.ogner@uni-mb.si, kacic@uni-mb.si

Abstract

The paper presents the speaker normalization technique we implemented in a teaching and training system for hearing handicapped children with the goal to reduce inter-speaker variability in time-frequency speech representation. In an effort to reduce variance caused by variation in vocal tract shape among speakers, a formant based nonlinear frequency warping approach to vocal tract normalization is investigated.

The proposed method can be efficiently realized in an Analysis by Synthesis framework. After the speech decomposition into the vocal tract envelope and excitation model, the vocal tract envelope is warped by the estimated frequency warping function, while the excitation characteristics are mapped to the reference speaker excitation.

The results have shown significant spectral distance decrease for correctly pronounced words between test and the reference speaker after the normalization has been applied, while for poor pronunciation by the test speaker the spectral distance remains relatively high.

1. Introduction

The main source of inter-speaker variability in time-frequency representation of speech is due to the differences in vocal tract length and in certain degree variability in vocal excitation characteristics. Both components appear at somewhat different scale in its spectrum, the fine scale structure is mostly due to the excitation, while the intermediate scale structure is due to the vocal tract transfer function, characterized by its resonant frequencies. When only the visual channel is available, as in the case of hearing handicapped persons, accurate evaluation of correct articulation is almost impossible, since different speakers have different formant frequencies for the same phoneme, hence visually perceived as local maximums positioned at different locations in time-frequency plane. Therefore some effective vocal tract normalization is necessary.

A common technique to suppress the influence of the excitation on spectrum variance is to smooth the spectrum by linear convolution or cepstral smoothing and preserving only the spectral envelope. For higher pitched speech, such as children’s speech, the accurate estimation of spectral envelope is a difficult task as the spectrum is less frequently sampled and vocal excitation often interferes with the spectral envelope estimation, especially at lower frequencies. Nevertheless, by suppressing the excitation component important information about voicing, necessary at the vocal tract normalization stage, is lost, since it does not make sense to calculate formant frequencies based on unvoiced speech data. Preserving the excitation information in the spectrum requires additional excitation spectrum normalization, since variability in the fine scale structure in the spectrum, e.g. pitch, causes variability in energy distribution among frequency bands.

In order to perform the vocal tract length normalization and excitation spectrum normalization simultaneously, the appropriate speech model is needed that would enable us the spectrum decomposition into spectral envelope and excitation model. The model has to be capable of spectrum resynthesis after some sort of manipulation on both components. Analysis by synthesis based on sinusoidal speech model offers such a speech signal parameterization appropriate for modifying speaker related characteristics, while preserving important phonetically relevant information in speech signal.

The paper is organized as follows. Section 2 gives the brief review of Harmonic+Noise sinusoidal speech model commonly used for speech coding. The attention is concentrated on details related to accuracy of spectrum decomposition into the source and spectral envelope estimation. Section 3 proposes the vocal tract normalization technique developed from vocal tract envelope warping by applying the nonlinear frequency warping function extracted from the test and reference speaker formant patterns. Section 4 provides method for excitation spectrum normalization based on the Harmonic+Noise excitation model. Section 5 describes the objective tests employed for the evaluation of the technique and presents the results.

2. Spectrum decomposition

Speech vocoders that utilize speech production model are very suitable for speech modification in speech synthesis and consequently for speaker modifications. Harmonic+Noise speech model decomposes input speech into vocal tract envelope and excitation model, hence it fits well in to the jointly vocal tract and excitation normalization scheme. Discussed implementation is one of the many realizations, probably closest to the MBE (Multiband Excitation) [1] speech coder with some modifications.
A major assumption made in this paper is that vocal tract length of the speaker is a long-term speaker characteristic. Therefore, when sufficient amount of observations drawn from production of different sounds is available, we can expect that estimated warping function truly characterizes vocal tract. The proposed method does not estimate VTL itself, it does map the test speaker vocal tract resonance pattern to some average or reference speaker resonance pattern observing the formant frequencies. The frequency warping function estimation is then reduced to model estimation that
best fits observed values. The constructed data pairs \((x_k, y_k)\) consists of measured values from previous section where \(x_k\) and \(y_k\) indicates observed formant frequencies obtained from reference and test speaker respectively, where \(i = 1 \ldots N\). The data are binned into \(k\) \(\{1 \ldots K\}\) categories, each corresponding to \(k\)th formant. Such a scheme can be best explained with the picture 1. Each of the data points indicates that at the frequency \(x_j\) on the reference frequency scale the warped test frequency scale should point to the frequency \(y_j\). The problem somewhat reminds on DTW. Estimate of the model parameters is then obtained by minimizing the chi-square quantity

\[
\chi^2 = \sum_{i=1}^{N} \left[ \frac{y_k - \sum_{j=1}^{J} a_j X_j(x_k)}{\sigma^2_i} \right]^2.
\]

(9)

Minimizing \(\chi^2\) leads to the linear least squares even that basis functions \(X_j(x)\) can be nonlinear. In matrix notation the solution is

\[
[A^T \cdot A] \cdot a = A^T \cdot b,
\]

(10)

where \(A\) is design matrix \(A_{ij} = X_j(x_i)\), \(b\) is vector defined by \(b_i = y_k\) and \(a\) is a solution vector. The experiments showed that the set of basis functions \(X_j(x) \in \{1, x, x^2, \ldots, x^{J-1}\}\) (polynomial fit) provides good results. We have no intention to suggest the reader that such a choice has physical background in speech production theory. Since \(A^T A\) is positive definite, Cholesky decomposition is the most efficient way to solve the normal equation (10).

### 3. Excitation spectrum normalization

In the joint time-frequency visual representation of speech the presentation of the phonetically relevant futures is of main importance, that is presenting the slow time-varying vocal tract transfer function estimated on a short-time basis. The common techniques such as cepstral smoothing or time-frequency filtering [3] obtain the smooth envelope of the power spectrum and suppress the excitation. These techniques suffer for accuracy when child speech is analyzed. On the other hand reducing the spectrum resolution to the filter-bank energy FBE, the excitation is lost, but the energy distribution among critical frequency bands is still excitation dependent, especially at lower frequencies where the frequency bands are denser.

As the speech corrector system uses critical filter-bank analysis as a preprocessor, we decided to preserve the spectrum with all its harmonic structure and rather normalize the excitation. The normalization of the excitation spectrum is carried out by first computing the excitation model parameters for both reference and test speaker:

- The voiced part of the excitation is synthesized by taking pitch from the reference speaker.
- The unvoiced (noisy) part of the excitation is synthesized with the same realization of the periodic random noise used for generating the excitation for reference speaker.
- Both components are mixed together according to the frequency dependent mixture function estimated for the test speaker.

### 5. Experiments

It’s difficult to evaluate the efficiency of the method because of its specificity. We compared spectral distance between test and reference speaker before and after the normalization and for poor and for excellent pronunciation from the test speaker. In our experiments we expected that the distance for correct pronunciations after the normalization would be as small as possible and for poor articulation would stay practically unchanged. The database has been comprised of thirty speakers: male, female and children, ten speakers from each group. The words used were Hungarian isolated digits.

<table>
<thead>
<tr>
<th>test/reference</th>
<th>word</th>
<th>correct [%]</th>
<th>incorrect [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>male/child</td>
<td>/0/</td>
<td>55.9</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>/1/</td>
<td>55.4</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>/2/</td>
<td>38.8</td>
<td>11.5</td>
</tr>
<tr>
<td></td>
<td>/3/</td>
<td>46.6</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>/4/</td>
<td>52.1</td>
<td>6.1</td>
</tr>
<tr>
<td>female/child</td>
<td>/0/</td>
<td>49.2</td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td>/1/</td>
<td>44.3</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>/2/</td>
<td>34.4</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td>/3/</td>
<td>35.8</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td>/4/</td>
<td>40.0</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Table 1: Weighted spectral decrease after the normalization for correct and incorrect pronunciation.
The table 1 shows the averaged spectral distance decrease expressed in percents after the normalization takes place. As it can be seen, the decrease is approximately 50% in the case of male to child and 40% in the female to child case.

Because we didn’t have proper database of speech handicapped children, the poor pronunciation has been simulated with incorrect pronunciation by inserting the arbitrarily chosen wrong words instead of poor test speaker articulation. The reported values are absolute, because the distances were sometimes higher and sometimes lower after the normalization and are about 10%.

### 6. Summary

In this paper we presented the spectrum normalization technique suitable for uniform visual presentations of speech characteristics for audio-visual articulation training systems. The spectrum was decomposed using Harmonic+Noise model and resynthesized after excitation spectrum normalization and spectral envelope warping. The experiments were conducted by spectral distance comparison before and after the normalization. The results showed the significant distance decrease after the normalization while for the incorrect pronunciation the spectral distance remained in acceptable limits.

### 7. References


---

Picture 2: Spectrograms for the word /nula/; frequency resolution 20 bands. From top to bottom: male test speaker, reference child speaker; test speaker with VTL normalization; test speaker with VTL and excitation normalization.